**Светла Коева**
(София)

# КОМПЮТЪРНА ЛЕКСИКОГРАФИЯ – ТРАДИЦИИ И ПРЕДИЗВИКАТЕЛСТВА[1]

**Abstract**: The paper presents some definitions for a corpus and the main features of modern corpora. It is shown that the computational lexicography has its own traditions in Bulgaria. At the same time not full potential of computational linguistics approaches and methods is in use.

**Key words**: computational lexicography, corpus

Накратко ще бъдат разгледани основният обект и методите на изследване, които се прилагат в корпусната лингвистика, компютърната лексикография и компютърната лингвистика. И в трите изследователски области корпусите намират приложение. В корпусната лингвистика корпусите са основен обект на изследване, лингвистични заключения се правят на базата на реалната употреба на езиковите единици, която се наблюдава в корпусите. В компютърната лексикография корпусите представляват източник за извличане на различен тип лингвистична информация (със средства, заимствани от компютърната лингвистика), която, най-общо казано, подпомага лексикографската работа. В компютърната лингвистика корпусите се използват за създаване на езикови и преводни модели, за извличане на различен тип информация и надеждни езикови данни. Следователно корпусите се използват и в трите научни области с относително сходно предназначение: за получаване на необходимите данни за лингвистичен анализ и езикови обобщения. Корпусите, в това число и Българският национален корпус (http://www.ibl.bas.bg/BGNC_bg.htm), представляват надежден източник за наблюдение, анализ и изводи (подкрепени от обективни количествени и дистрибутивни данни) за синтаксиса и семантиката на различни езикови явления (за употребата им в различни стилове, жанрове и тематични области), както и за автоматично извличане на езикови данни, езикови отношения и модели.

Известни са различни дефиниции за корпус, в които се подчертава значението на корпусите за лингвистичните изследвания: „*съвкупност от автентични езикови данни, които могат да се използват за лингвистични изследвания*“ (Лиич 1997: 1); „*колекция от езикови примери, които са избрани и подредени според експлицитни лингвистични критерии, за да се използват като модел на езика*“ (ЕГСКОЕ 1996: 4); „*съвкупност от езиков материал в електронна форма, подходящ за компютърна обработка с приложение в лингвистичните изследвания или езиковите технологии*“ (Лиич 1997: 1);

„*голяма колекция от езикови примери, представени по начин, даващ възможност за компютърна обработка, и избрани по определени (лингвистични) критерии, така че да представляват адекватен езиков модел*“ (Коева 2010: 9).

Съвременните корпуси съдържат милиарди думи и анализът на данните в тях може да се извърши ефективно и надеждно. Вече традиционно се приема, че по-големите корпуси представят по-достоверна информация за езика, тъй като се предполага, че по-голямото количество данни е предпоставка за илюстрирането на по-широк кръг от езикови явления, при това с по-надеждна информация за честотата на срещане и дистрибуцията в различни стилове, жанрове и тематични области. Необходимостта от обективи данни, от една страна, и все по-голямото количество от достъпни електронни текстове, от друга страна, са естествени предпоставки за това големината на корпусите да се увеличава непрекъснато. Неслучайно в момента интернет е един от основните източници за събиране на корпусни единици (Килгариф, Грефенщет 2003). Представителността и балансираността като характеристики на корпуса са насочени към това да гарантират доколко даден корпус (съответно съотношението между текстовете, които са включени в него) може да предостави адекватни факти за езиковата действителност към даден момент. Известни са много опити двете характеристики да се дефинират по един или друг начин (Аткинс 1992/1993; Байбър 1993; Лиич 2005; Синклер 2005; МакЕнери и др. 2006), но до момента няма ясни критерии, които да осигуряват последователното им прилагане при изграждането на произволен корпус. При създаването на корпуси се наблюдават два основни метода. Единият предполага създаване на модел за структурата на корпуса и последователно запълване на тази структура със съответните текстове. Този подход е характерен за началото на създаването на корпуси (Франсис, Кучера 1964) и по обективни причини все по-малко се следва поради: а) невъзможност да се създаде модел, който да е достатъчно адекватен, като удовлетворява различни критерии; б) трудно запълване на модела с подходящи текстове особено ако се поставят по-големи изисквания по отношения на обема на корпуса. Вторият метод, ако може да се нарече така, се основава на събиране на текстове без наличието на предварително предпоставен модел за структурата на корпуса. В момента подобен начин за създаване на корпуси получава все по-широко разпространение особено като се има предвид тезата, че по-големият обем предполага и по-добра възможност за илюстрация на езиковите явления, следователно големият обем в значителна степен обезсмисля наличието на предварително дефиниран модел за структурата на корпуса. В тази светлина представителността на даден корпус може да се разглежда като съвкупност от два фактора: възможно най-голям обем и възможно най-голямо разнообразие на текстовете по отношение на техния стил, жанр и тематична област. Балансираността като характеристика противоречи на изискването за обем и представителност (разбирана като многообразие) – големият обем и стиловото и жанровото

многообразие трудно се съвместяват с равномерното застъпване на предварително дефинирани категории (а и всяка субективна намеса в крайна сметка може да изкриви анализа). За конкретни лингвистични задачи може да се дефинира пропорционално разпределение на текстове на базата на съществуващото многообразие от езиков материал. В този случай е по-добре да се говори не за балансираност, а за структура, която се подчинява на определени критерии, които се предполагат от поставянето и решаването на конкретна задача. Извличането на корпуси с предварително дефинирана структура може да се направи успешно, ако корпусните единици са описани с разнообразни метаданни (автор, стил, жанр, тематична област, година на издаване, брой думи и т.н.) – колкото е по-подробно и прецизно описанието, толкова са по-големи и разнообразни възможностите за извличане на различни по своята структура подкорпуси. По този начин се избягва необходимостта да се следва предварително зададен модел и да се търсят корпусни единици, които съответстват на модела.

Българският национален корпус (Коева и др. 2012) е голям (според съвременното разбиране) динамично развиващ се корпус – разширява се с нови текстове, които се класифицират с помощта на подробни метаданни. Дизайнът на корпуса осигурява поддръжка на едноезикови и многоезикови корпуси посредством унифициран подход за тяхната компилация, предварителна обработка, анотация и достъп. При описанието на корпусните единици се цели детайлност, която позволява лесна класификация на текстовете по различни критерии и преструктуриране въз основа на различни класификационни признаци. Българските текстове в Българския национален корпус наброяват 1.2 милиарда думи, които са разпределени в над 240 000 текстови документа (към края на януари 2013 г.). Оригиналните текстове на български език съставляват 37.1% от всички текстове, преводните – 40.5%, а за останалите 22.4% няма информация. Преобладаващото количество текстове са писмени документи – 97.35%. По-голяма част от текстовете – 97.6%, са събрани от интернет (посредством автоматично – основно, или ръчно изтегляне). Към момента задачата за разширяване на Българския национален корпус се свежда до автоматично идентифициране в интернет на подходящи текстове, автоматично изтегляне на документите, автоматично извличане на детайлни метаданни и автоматична лингвистична анотация.

Начало на модерната компютърна лексикография се свързва с проекта COBUILD (Синклер 1987). В рамките на проекта се създава речниково описание за английски език, което отразява актуалната употреба на думите в голям и разнообразен корпус от текстове. Корпусната лингвистика намалява в значителна степен субективността на традиционните подходи при дефиниране на значенията, като изследва езикови данни за множество срещания на дадена лексикална единица в големи по обем корпуси (Синклер 1987), а значението на думите в речниците се описва след анализ и синтез на информация за употребата в контекст (Аткинс 1992/1993). В етапа на анализ се търсят колкото

може повече обективни лингвистични факти за значенията на дадена дума (Аткинс 1992/1993: 33) чрез изследване на употребата на думата в корпуси, в съществуващи архиви и речници. Данните се разделят така, че лексикалните единици в дадена група от примери да имат общо значение и да се различават от значенията в останалите групи, като за всяка група се определя какво я обединява. В етапа на синтез направените заключения се обобщават в ограниченията на езика на тълковните дефиниции. Отбелязва се, че системното приложение на корпусно базираната методология отразява парадигматична промяна в съвременната лексикография (Рындъл, Килгариф 2011).

Компютърната лексикография в наши дни далеч надхвърля прилагането на примери, които са извлечени от реалната им употреба в корпуси. Съвременната лексикография си служи успешно с корпусите при решаването на редица задачи като: извличане и подбор на словник за конкретен тип речник; анализ, базиран на корпусни данни; идентифициране на значенията на лексикалните единици и формулиране на подходящи дефиниции; определяне на основните признаци на лексикалните единици: съчетаемост, синтактично поведение, употреба в различни тематични област и разбира се – извличане и подбор на подходящи примери. Както се посочва (Благоева, Колковска 2011: 17–18), корпусни и традиционни методи за създаване на словник се използват при работата върху последните няколко тома на многотомния тълковен Речник на българския език (РБЕ 2008; РБЕ 2012) – словникът се изработва на базата на съществуващите лексикални картотеки и речници на българския език и се допълва с автоматично генерирани списъци от думи на базата на Българския национален корпус. Българският национален корпус се използва и при изготвянето на словник за Речника на новите думи в българския език (Пернишка, Благоева, Колковска 2010). Сравнени са два субкорпуса от Българския национален корпус, съдържащи текстове, публикувани от 1945 до 1989 г. и след 1990 г., в резултат на което е извлечен честотен списък от словоформи, регистрирани само в по-новите текстове (Благоева 2009). Отчита се не само фреквентността, но и броят на текстовете, в които са регистрирани срещанията, т.е. използва се информация и за дистрибуцията на новите думи. От доста време не е проблем да се съставят честотни списъци не само на словоформи, но и на основни форми – така броят на срещанията на думите е по-адекватно отразен. На базата на честота на срещания на основни форми и дистрибуцията им се получава достоверна информация (подкрепена със статистически данни) за употребата за лексикалните единици с висока, средна и ниска честота (нови думи, редки и остарели думи и т.н.). Могат да се прилагат филтри (прости граматики), които да извадят списъци от думи, които имат или нямат множествено число, имат или нямат някои граматични конструкции и т.н.

Известен похват в корпусната лингвистика е извличането на редове от конкорданси за дадени лексикални единици, което улеснява наблюдението върху различните употреби посредством сортиране на ляв и десен контекст с определена дължина – по този начин се подпомага разграничаването на от-

делните значения и описанието на синтактичните и семантичните свойства на лексикалните единици. Лексикалната съчетаемост се илюстрира посредством извличане на колокации – редовно срещащи се словосъчетания с дадена дума, които могат да бъдат както съставни думи, така и свободни словосъчетания. Скиците на употреба на думите представляват корпусно базирано обобщение за граматичните и колокационните свойства на думите (Килгариф, Тъгуел 2001).

Познати са различни подходи при представяне на значението на съставните думи – включва се към значението на главната дума, съотнася се с пообща дефиниция, която е вярна за множество съставни думи, представя се като отделна лексикална единица със собствени значения. Практика в (българските) речници е съставните думи да се изброяват (непоследователно) към речниковата статия на една от думите в състава си, без винаги това да е главната дума. Съставните думи би трябвало да се интерпретират равнопоставено с останалите думи, ако изразяват уникално понятие – като тест за разпознаването им може да се използва тестът за еквивалентност с дума в същия или друг език и възможността за субституция с тази дума или с нейните хипероними (Коева 2006). Както в повечето области на компютърната лингвистика, автоматичното извличане на кандидати за съставни думи от корпуси (различни от свободните словосъчетания) се основава на три различни стратегии: лингвистични техники, статистически методи и комбинирани подходи. Лингвистичните техники използват информация за морфологичната и синтактичната структура на съставните думи и са различни за различните езици. Например следните последователности от категории (за които се очаква, че образуват именна група) са с честота на срещане над 10% в текстове с правни документи на Европейския съюз:

AN → *риболовен сезон, земеделски цени, термална енергия, климатична инсталация*

NpN → *обогатяване на гориво, подобряване на почвата, права на детето, свобода на печата*

NpAN → *опазване на околната среда, номенклатура на земеделските продукти, използване на слънчевата енергия, средства за масова информация*

AAN → *семейно земеделско стопанство, европейска парична система, интелигентна транспортна система, магнитен информационен носител*

ANpN → *електронен трансфер на фондове, оптическо разпознаване на символи, правна уредба на телекомуникациите, избирателно разпространение на информацията*

Следвайки честотния анализ на конституентната структура на българските съставни думи, синтактичните модели, които са обект на анализ, се разпознават с помощта на подходящи синтактични шаблони. Статистически-

те техники, от своя страна, се основават на различните статистически свойства на съставните думи по отношение на останалата лексика и се базират на откриването на последователности от думи с по-голяма честота над някакъв определен праг. Обикновено целта е да се измери свързаността (асоциацията) между частите на съставните думи, като се провери дали частите им се срещат заедно случайно, или не. Статистическите методи приписват числена стойност на кандидатите за съставни думи, за да ги потвърдят или изключат. Измерването на лексикалната свързаност съотнася наблюдаваната с очакваната честота. Един от най-често използваните коефициенти е Log-likelihood (Дюнинг 1993), тъй като се приема, че има най-добри резултати сред другите подобни изчислителни методи. Колкото е по-голяма стойността на Log-likelihood, толкова е по-силна асоциативната свързаност между двете думи, следователно последователността от думи е по-вероятен кандидат за съставна дума. Добре известно е, че едно и също понятие може да се изрази по различен начин и автоматичното извличане на съставни думи трябва да може да разпознава и свързва различните лингвистични форми: ортографски (главни и малки букви), флективни (словоформите), словообразователни (деривати), синтактични (словоред) и семантични (синоними). Честотните стойности могат да се преизчислят според групирането на словоформите към една и съща лема. Така морфологично свързаните срещания на дадена съставна дума се съотнасят и се разглеждат като една и съща дума.

Използват се и по-сложни методи. Например, изхождайки от предпоставката, че свободните словосъчетания допускат синонимна замяна на компонентите си, а колокациите (разбирани в смисъл на съставни думи) – не, се разработва метод за извличане на колокации, като се измерва семантичната свързаност – до каква степен даден кандидат за съставна дума допуска замяна със синоними на елементите си (Пиърс 2001).

Извлечени от големи по обем корпуси, данните за фреквентност позволяват адекватна оценка на съвременната употреба на лексиката и нейната дистрибуция по жанрове и стилове и тематични области. Следваща крачка е да се разграничи честотата на срещане на различните значения, с които дадена дума е употребена в текста – това може да стане, като се изчислят близки или еднакви контексти на употреба на синтактично и семантично равнище в текстове, принадлежащи към различна тематична област, жанр или стил, които също са релевантни за употребата на едно или друго значение. В най-простия случай примерите за употреба в корпус попадат в една или повече различни групи и всяка група, ако е достатъчно голяма и достатъчно различна от останалите групи, оформя ново значение (Килгариф 1997: 108) – значенията трябва да се разбират като абстракции по отношение на групи от примери. Друг начин за автоматично диференциране на значенията на дадена дума използва следната предпоставка: ако дума се превежда по няколко семантично различими начина на друг език, то това е доказателство за различно значение (Браун и др. 1991; Гейл и др. 1992). Съвременните технологии предлагат

много начини за извличане на списъци на възможните значения, с тежести в зависимост от вероятността им; на срещанията на всяко значение в корпуси; на общи субкатегоризационни фреймове, с тежести в зависимост от вероятността им; на общи селективни ограничения, с тежести в зависимост от вероятността им, и т.н.

Не на последно място, използването на специализирани системи за създаване на речници позволява процесът на създаването на речник да бъде подчинен на концепцията за създаване на речника – веднъж зададена, концепцията за структурата на речника не може да бъде променяна неволно или по грешка, тъй като се контролира от системата за създаване на речника. Лексикографските данни, които принадлежат към затворени класове, се избират от предварително зададени списъци и по този начин възможността за грешка се ограничава още повече. Възможността да се правят справки, някои връзки да се създават и/или проверяват автоматично допълнително улеснява лексикографската работа.

Езиковите корпуси намират широко приложение в съвременната лингвистика като източник за извличане на езикови данни, въз основа на които се изграждат и проверяват различни лингвистични хипотези. Все по-широкият кръг от приложения налага нови изисквания към корпусите, а оттам и нови принципи при тяхното създаване, структуриране, документация, анотиране и обработка:

– Съставяне на все по-големи по обем корпуси, отразяващи максимално достоверно състоянието на даден език през определен етап от неговото развитие.

– Все по-точно и детайлно автоматично анотиране на корпусите с подходяща лингвистична информация и метаданни.

– Все по-широко използване на компютърната лингвистика за обработка на лингвистичните корпуси и приложение на получените резултати в различни области, включително в лексикографията.

## БЕЛЕЖКИ

[1] Изследването е проведено с финансовата подкрепа на Фонд „Научни изследвания“, договор ДТК 02–53/2009.

## ЛИТЕРАТУРА

**Аткинс 1992/1993**: Atkins, B. T. S. Theoretical lexicography and its relation to dictionary-making. // *Dictionaries*. *Journal of the Dictionary Society of North America* 14, pp. 4–43.

**Байбър 1993**: Biber, D. Representativeness in corpus design. // *Literary and Linguistic Computing* 8 (4), pp. 243–258.

**Благоева 2009**: Blagoeva, D. Electronic Corpora and Bulgarian New-Word Lexicography. // *Études Cognitives*. Vol. 9. Warszawa, SOW, pp. 143–150.

**Благоева, Колковска 2011**: Д. Благоева, С. Колковска. Корпусният подход в българската лексикография – практика и перспективи. // *Съвременни методи и подходи в лексикографската практика.* Сборник студии и статии. София: Авангард Прима, с. 7–45.

**Браун и др. 1991**: Brown, P. F., S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer. Word-sense disambiguation using statistical methods. // *Proceedings of the 29th Conference of the Association for Computational Linguistics*, pp. 264–270.

**Гейл и др. 1992**: Gale, W., K. W. Church and D. Yarowsky. Using bilingual materials to develop word sense disambiguation methods. // *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp. 101–112.

**Дюнинг 1993**: Dunning, T. Accurate methods for the statistics of surprise and coincidence. // *Computational Linguistics*, 19 (1), pp. 61–74.

**ЕГСКОЕ 1996**: *Expert Advisory Group for Language Engineering Standards Preliminary recommendations on corpus typology*. EAG–TCWG–CTYP/P. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

**Килгариф 1997**: Kilgarriff, A. I don't believe in word senses. // *Computers and the Humanities* 31 (2), pp. 91–113.

**Килгариф, Грефенщет 2003**: Kilgarriff, A., G. Grefenstette. Introduction to the Special Issue on Web as Corpus. // *Computational Linguistics.* 29 (3), 2003, pp. 333–348.

**Килгариф, Тъгуел 2001**: Kilgarriff, A., D. Tugwell. WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. // *Proceeding ACL workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation*. Toulouse, pp. 32–38.

**Коева 2006**: Koeva, Sv. Inflection Morphology of Bulgarian Multiword Expressions. // *Computer Applications in Slavic Studies*. Sofia: Boyan Penev Publishing Center, pp. 201–216.

**Коева 2010**: Коева, Св. Българският семантично анотиран корпус – теоретични постановки. // *Българският семантично анотиран корпус*. Редактор и съставител Св. Коева. София, с. 7–43.

**Коева и др. 2012**: Koeva, Sv., I. Stoyanova, Sv. Leseva, Ts. Dimitrova, R. Dekova, E. Tarpomanova. The Bulgarian National Corpus: Theory and practice in corpus design. // *Journal of Language Modelling*, 1 (1), pp. 65–110.

**Лиич 1977**: Leech, G. Introducing corpus annotation. // Garside R., G. Leech, A. M. McEnery (eds.). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman, pp. 1–18.

**Лиич 2005**: Leech, G. Adding Linguistic Annotation. // Wynne, M. (ed.). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, pp. 17–29. <http://ahds.ac.uk/linguistic-corpora/>

**МакЕнери и др. 2006**: McEnery, T., R. Xiao, Y. Tono. *Corpus-based Language Studies. An Advanced Resource Book*. Routledge.

**Пернишка, Благоева, Колковска 2010**: Пернишка, Е., Д. Благоева, С. Колковска. *Речник на новите думи в българския език (от края на XX в. и първото десетилетие на XXI в.)*. София: Наука и изкуство. 515 с.

**Пиърс 2001**: Pearce, D. Synonymy in collocation extraction, WordNet and Other lexical resources: applications, extensions & customizations. // *NAACL 2001*. Pittsburgh: Carnegie Mellon University, pp. 41–46.

**РБЕ 2008**: *Речник на българския език.* Т. 13. София: АИ „Марин Дринов“, ЕМАС.

**РБЕ 2012**: *Речник на българския език.* Т. 14. София: АИ „Марин Дринов“.

**Рындъл, Килгариф 2011**: Рындъл, М., А. Килгариф. Автоматично създаване на речници: къде е границата? // *Български език*. № 3, с. 7–33.

**Синклер 1987**: Sinclair, J. M. (ed.). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.

**Синклер 2005**: Sinclair, J. Corpus and Text: Basic Principles. // Wynne, M. (ed.). *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, pp. 1–16. <http://ahds.ac.uk/linguistic-corpora/>

**Франсис, Кучера 1964**: Francis, W. N., H. Kucera. *Brown Corpus Manual.* Department of Linguistics, Brown University, USA. <http://icame.uib.no/brown/bcm.html>